Towards exploiting image correspondence for weakly supervised visual recognition

Sudipta N. Sinha
Microsoft Research

Microsoft Research Asia
Faculty Summit 2016
Intelligent and Invisible Computing

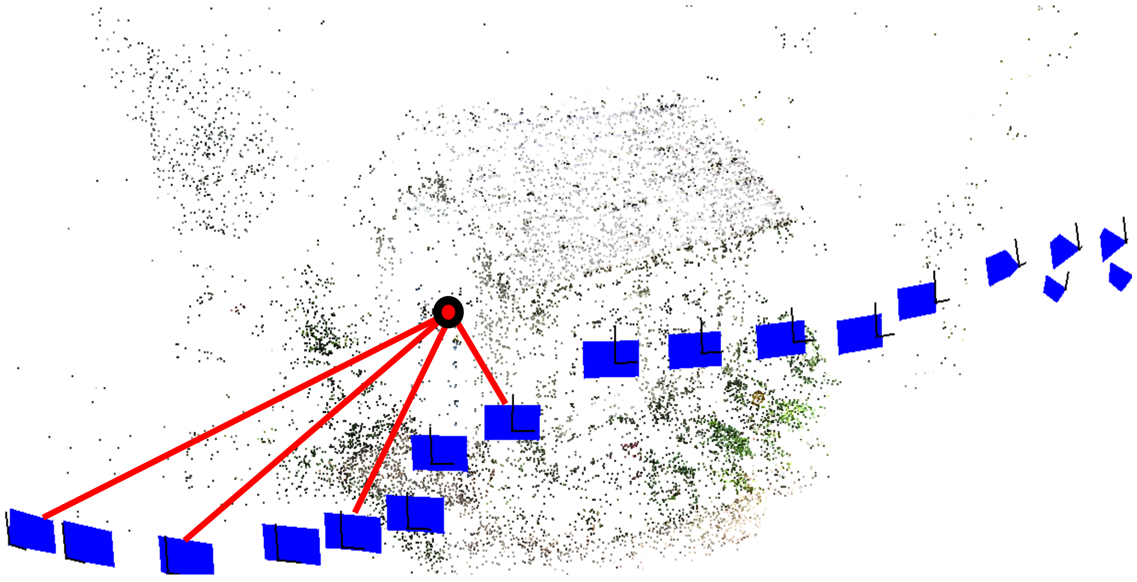
Microsoft

Introduction

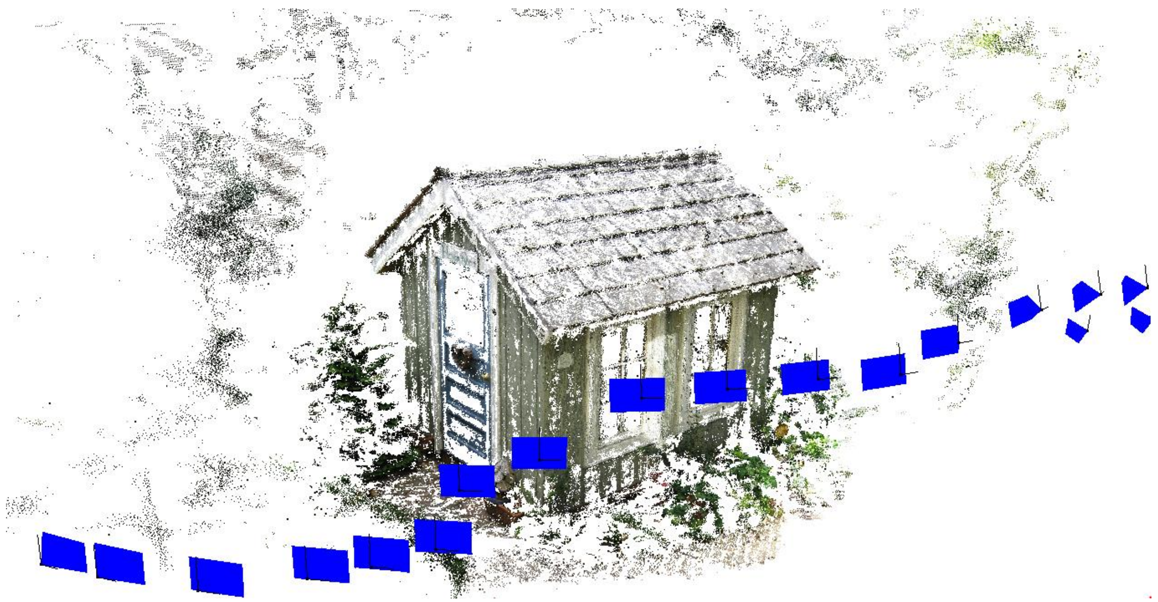
- **Machine learning in computer vision**
  - Major progress on discriminative tasks in supervised settings
    - Possible due to vast human-labeled image datasets
- **Collecting ground truth labels for images and video**
  - Mechanical Turk remains a major bottleneck
- **Correspondence problems in computer vision**
  - 3D scene reconstruction, image alignment
  - Source of indirect supervision
  - Open problems in unsupervised feature learning

Image Correspondence and 3D Scene Reconstruction

Sparse Pixel Correspondence



Sparse Structure from Motion (SfM)



Dense Reconstruction

Overview

- **Sparse Correspondence and Applications**
  - Place recognition
  - Color Transfer and Enhancing Photos
- **Dense Correspondence Estimation**
  - Stereo Matching on High Resolution Images and Video
- **Joint Correspondence and Cosegmentation**
  - Align images of different but semantically related objects

Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

[Bergamo, Sinha, Torresani, CVPR 2013]

"Tyn Church, Prague"

Input: Single image of tourist landmark.

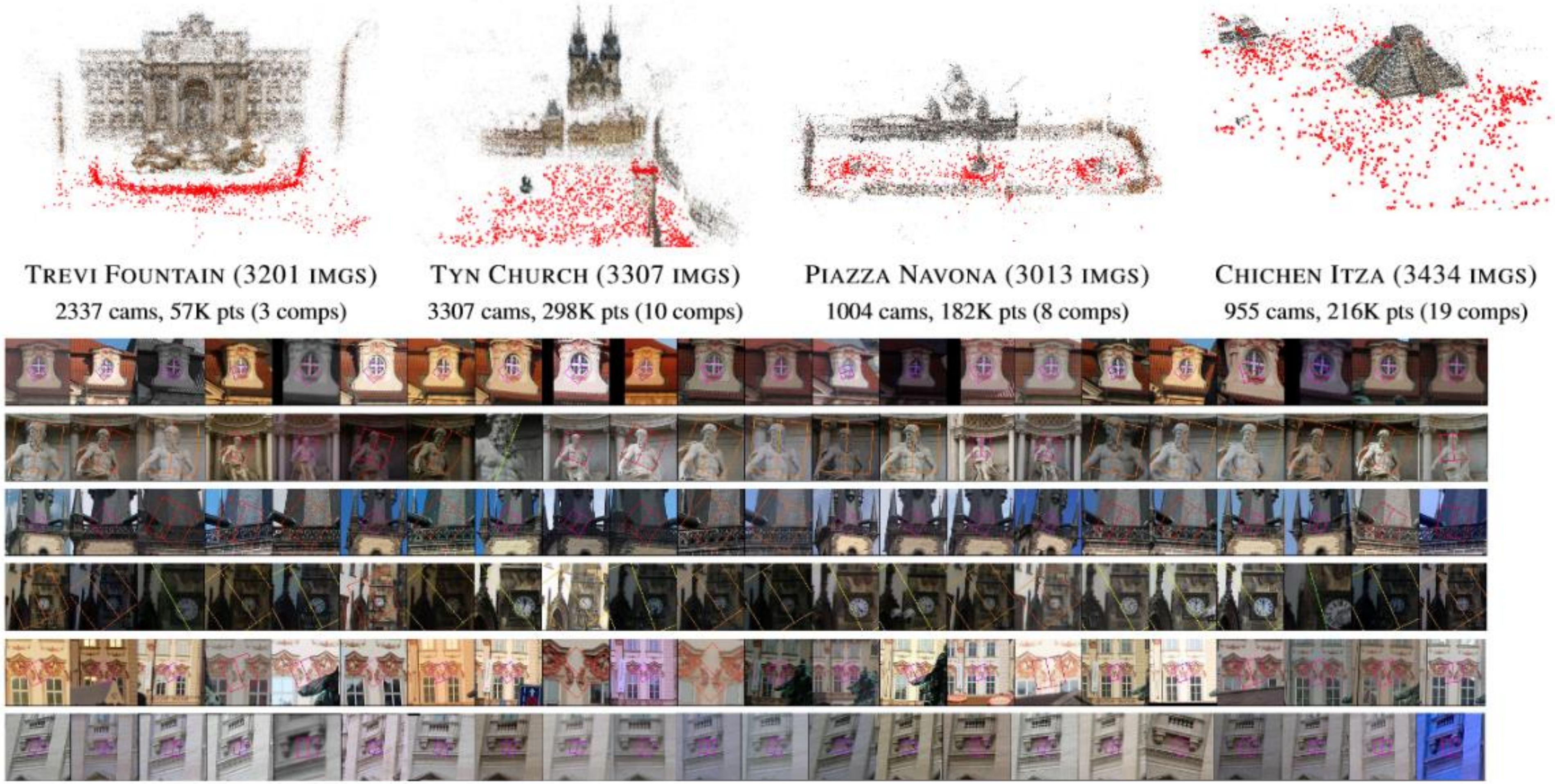
Task: Recognize the location.

Approach:
- Classification (instead of image retrieval)
  - one vs. all classifiers for each location

Main Idea:
- SfM reconstruction of landmark images (source: Flickr/Bing/Google…)
- Extract corresponding image patches from 3D SfM point cloud.
- Exploit correspondences to train discriminative features.

Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

[Bergamo, Sinha, Torresani, CVPR 2013]

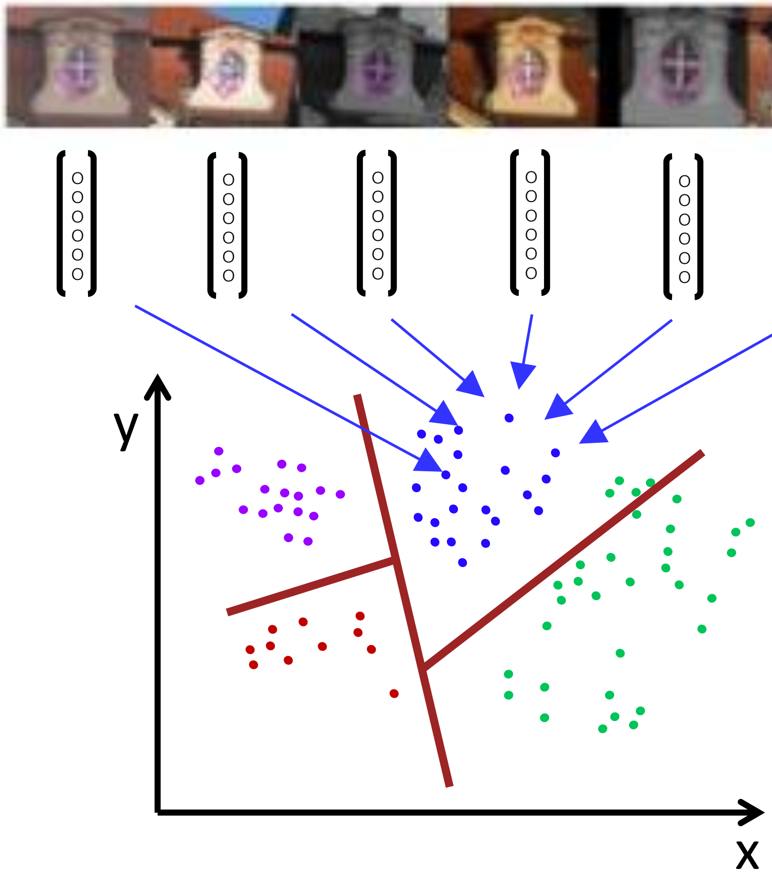


Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

[Bergamo, Sinha, Torresani, CVPR 2013]

- Random forest-based codebook
- Each track is a unique class
- Feature encoding
  - BoW / VLAD / Fisher Vector

- Track dataset:
  - Millions of tracks
  - Tens of millions of image patches



Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

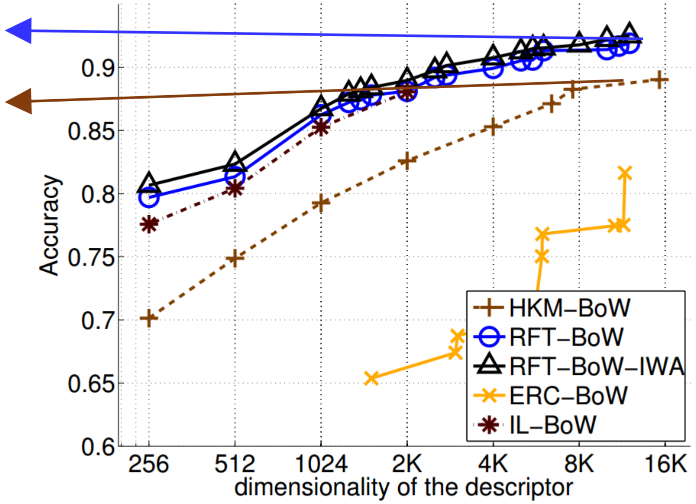
[Bergamo, Sinha, Torresani, CVPR 2013]

Ours: 92%

Baseline: 88%

Landmark3D

25 places,
5K test images

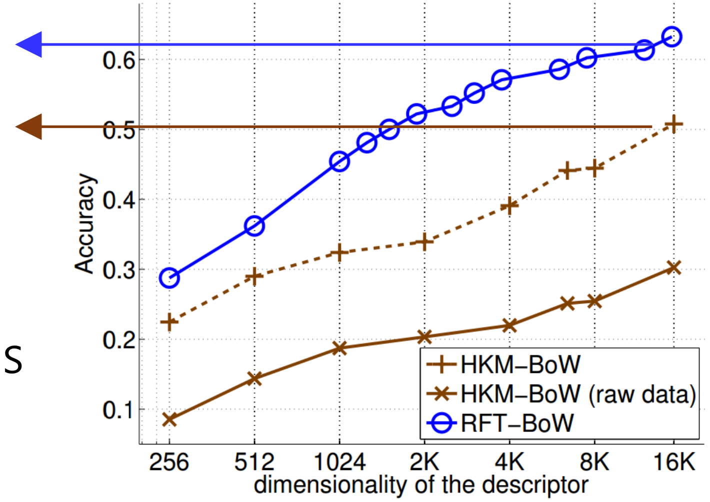


Ours: 65%

Baseline: 52%

Landmark-620

620 places
62K test images



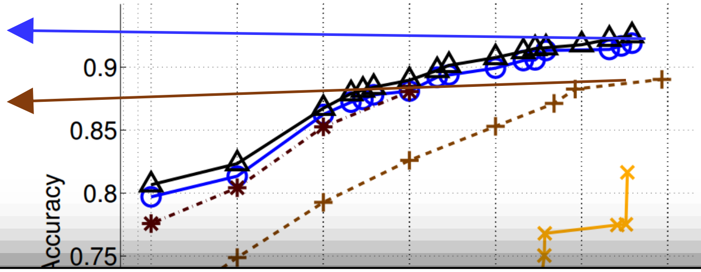
Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

[Bergamo, Sinha, Torresani, CVPR 2013]

Ours: 92%

Baseline: 88%

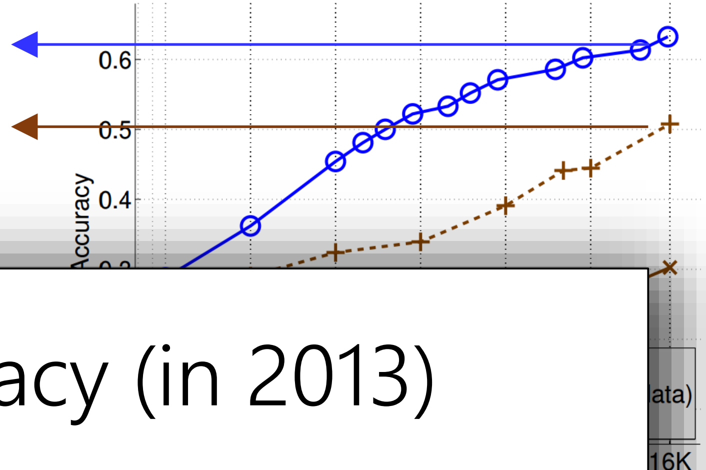
Landmark3D



Ours: 65%

Baseline: 52%

Landmark-620



- state of the art top-1 classifier accuracy (in 2013)
- outperformed unsupervised codebooks
- efficient codebook training and encoding schemes

Leveraging Structure from Motion to Learn Discriminative Codebooks for Landmark Classification

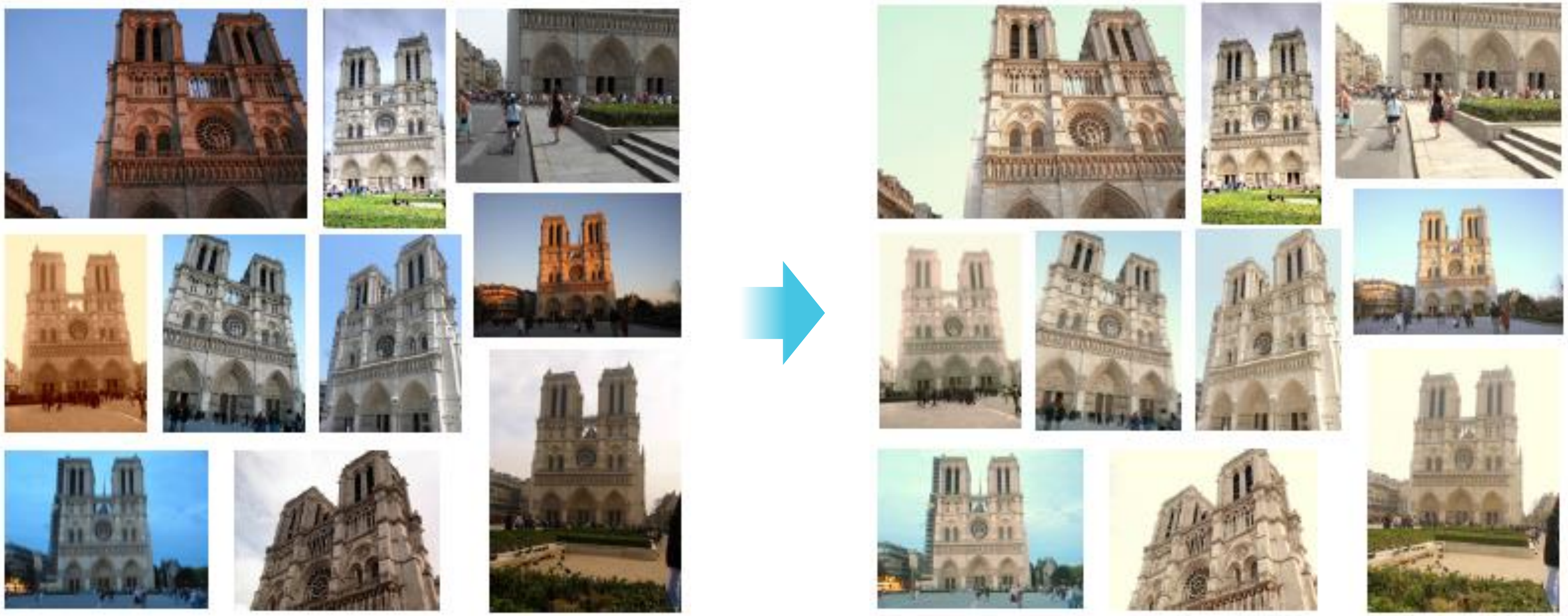
[Bergamo, Sinha, Torresani, CVPR 2013]

- **Almost automatic landmark classification system**
  - Image source: Internet photos (search engines, Flickr etc.)
- **Leveraged mature Structure from Motion (SfM) pipeline**
  - Filters outliers from Internet image collections.
  - Massive dataset of corresponding image patches.
- **Limitations**
  - Does not work for non-rigid scenes or objects
  - SfM only work on images of identical scenes

Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

Task 1: Improve color consistency of photos in a collection



Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

<u>Task 2:</u> Transfer the color of one photo to the rest in the collection

Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

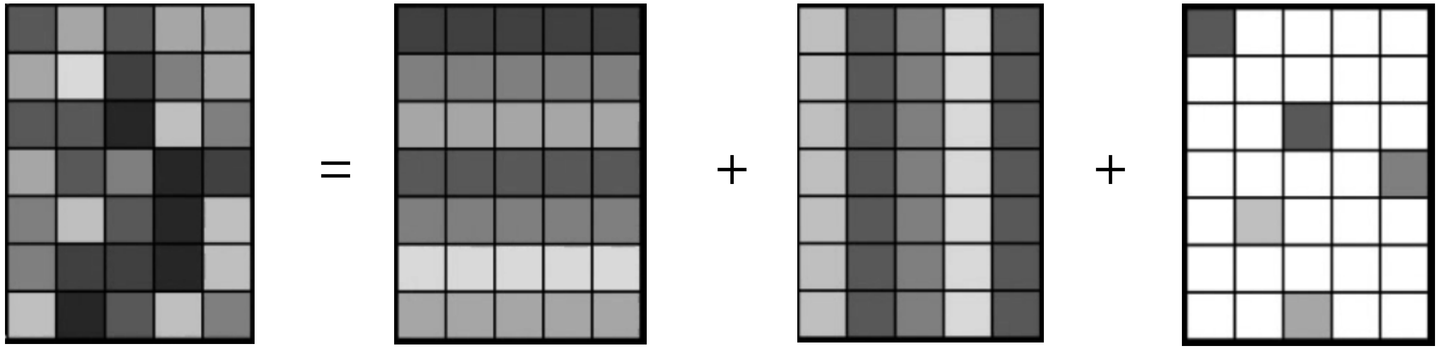
<u>Task 2:</u> Transfer the color of one photo to the rest in the collection

Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

Main Idea:

- Color Correction Model: $I' = (cI)^{\gamma}$
- Sparse image correspondences give constraints: $I_i(x_{ij}) = (c_i a_j e_{ij})^{\gamma_i}$
- Low-rank Matrix Factorization formulation



- Low-Rank Matrix Decomposition Technique [Cabral+ ICCV 2013]

Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

Application: Image Stitching

- Microsoft Research Image Composite Editor (ICE)
- Our correction makes the result more consistent

Photoshop CS6 correction

Using original images

Our correction

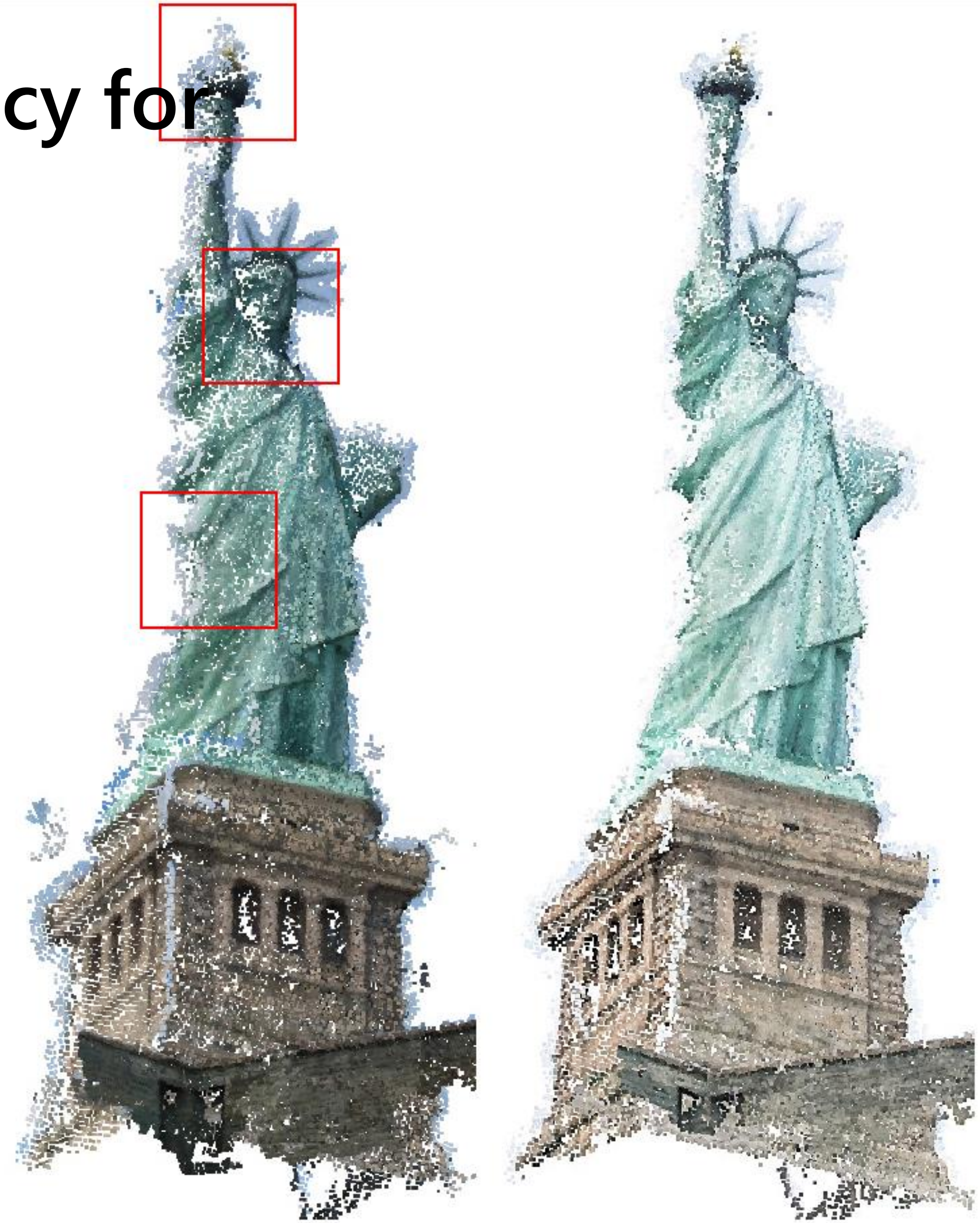
Efficient and Robust Color Consistency for Community Photo Collections

[Park, Tai, Sinha, Kweon, CVPR 2016]

Application: 3D reconstruction

- Corrected images produce a better 3D model

original images corrected images



Using original images Using corrected images

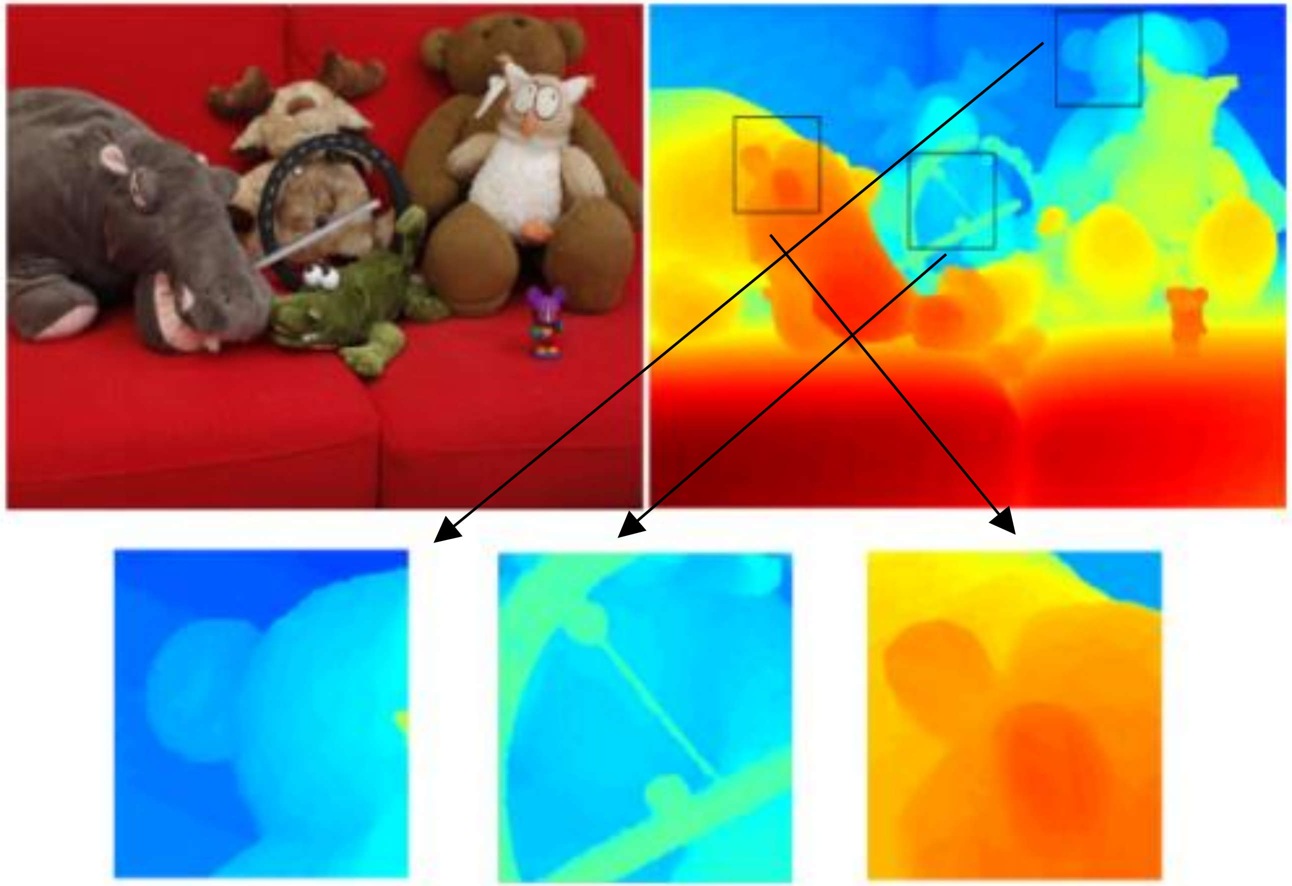
Overview

- **Sparse Correspondence and Applications**
  - Place recognition
  - Color Transfer and Enhancing Photos
- **Dense Correspondence Estimation**
  - Stereo Matching on High Resolution Images and Video
- **Joint Correspondence and Cosegmentation**
  - Align images of different but semantically related objects

Efficient High-Resolution Stereo Matching using Local Plane Sweeps

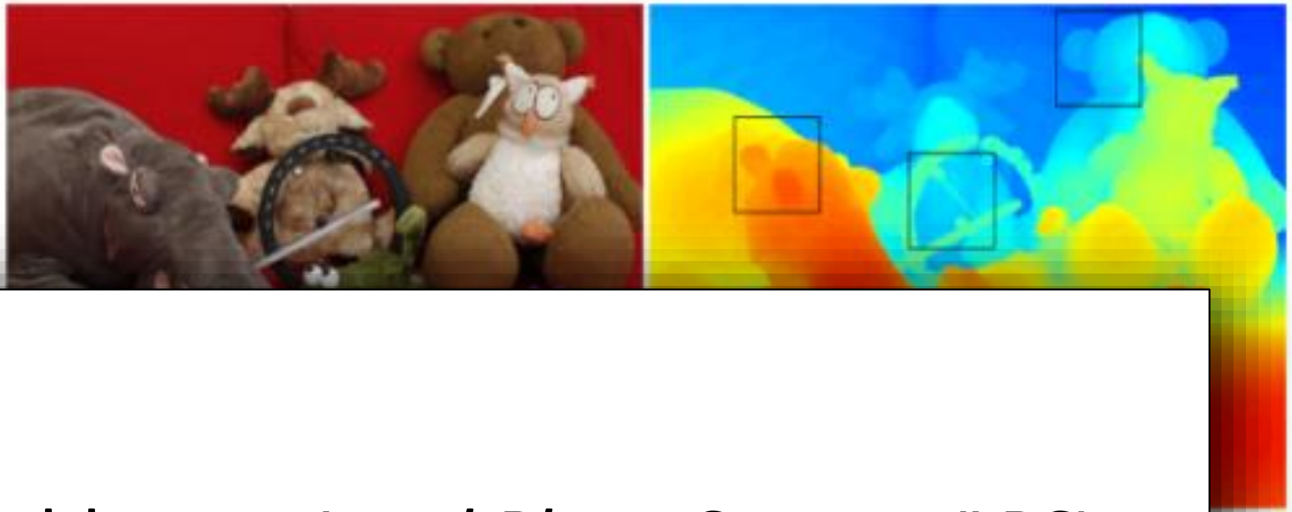
[Sinha, Scharstein, Szeliski, CVPR 2014]

- High Resolution
  - 10+ MPixels
- Large disparity range
- Global stereo methods
  - Evaluate all disparities
  - Impractical !



Efficient High-Resolution Stereo Matching using Local Plane Sweeps

[Sinha, Scharstein, Szeliski, CVPR 2014]



- High Resolution
  - 10+ MPixels

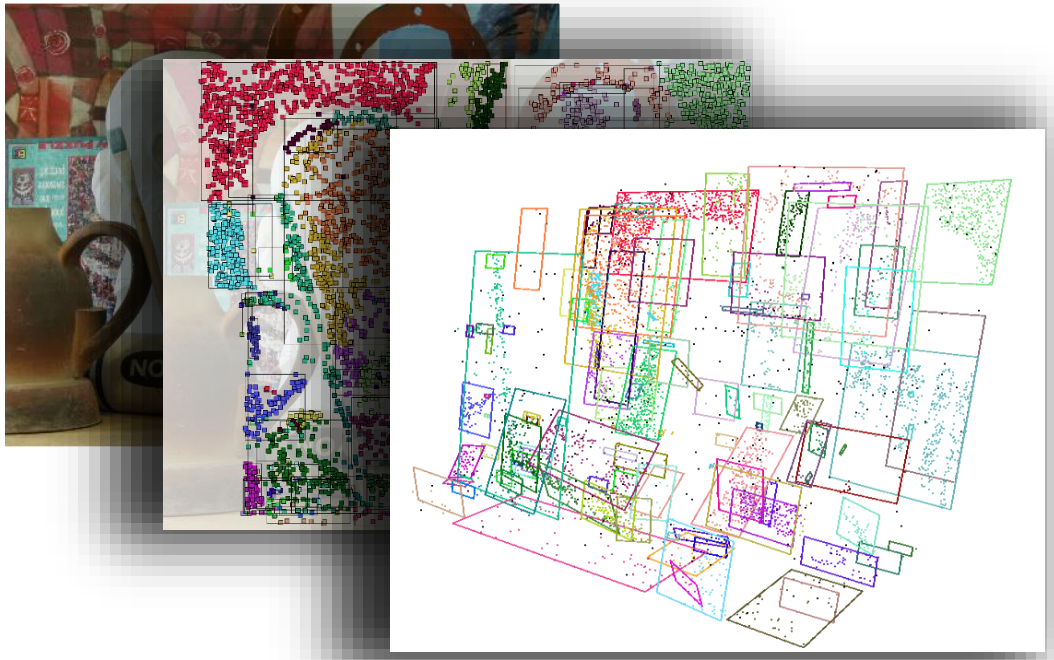
Main Idea

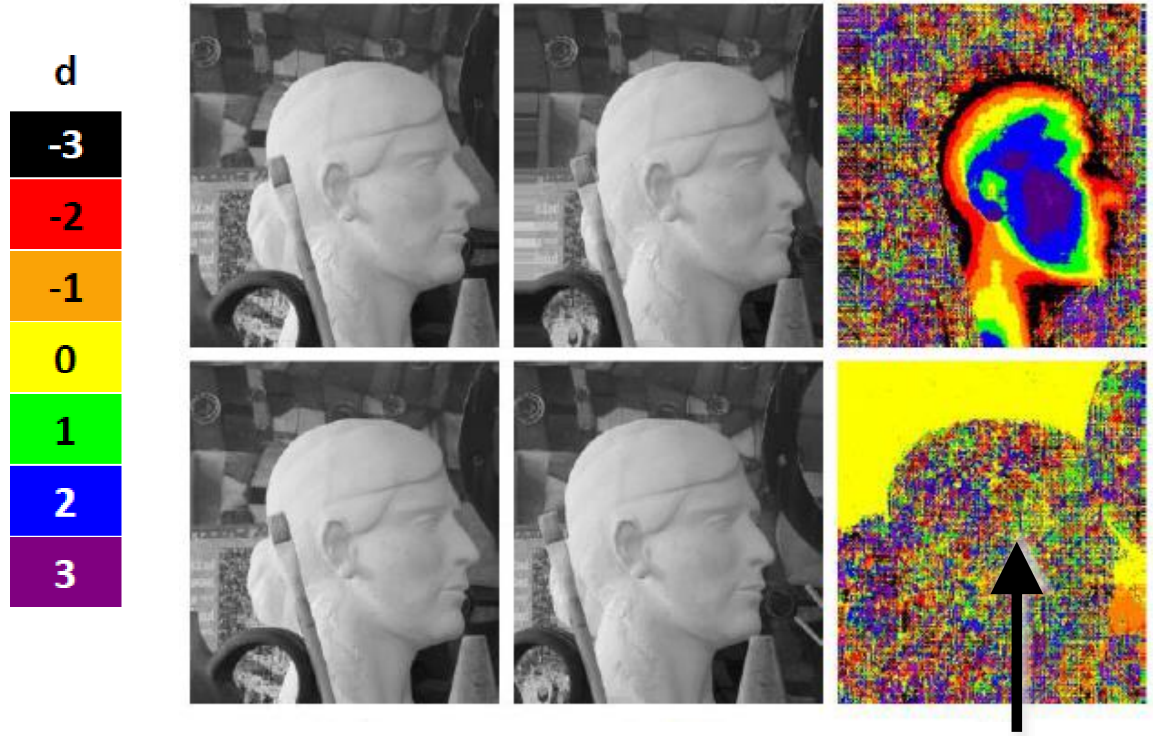
- Solve many local stereo problems – *Local Plane Sweeps (LPS)*
- Generates surface proposals.
- Fuse proposals to obtain disparity map

Efficient High-Resolution Stereo Matching using Local Plane Sweeps

[Sinha, Scharstein, Szeliski, CVPR 2014]

- Match sparse features, robustly identify planes
- Each local problem explores a fraction of the full search space

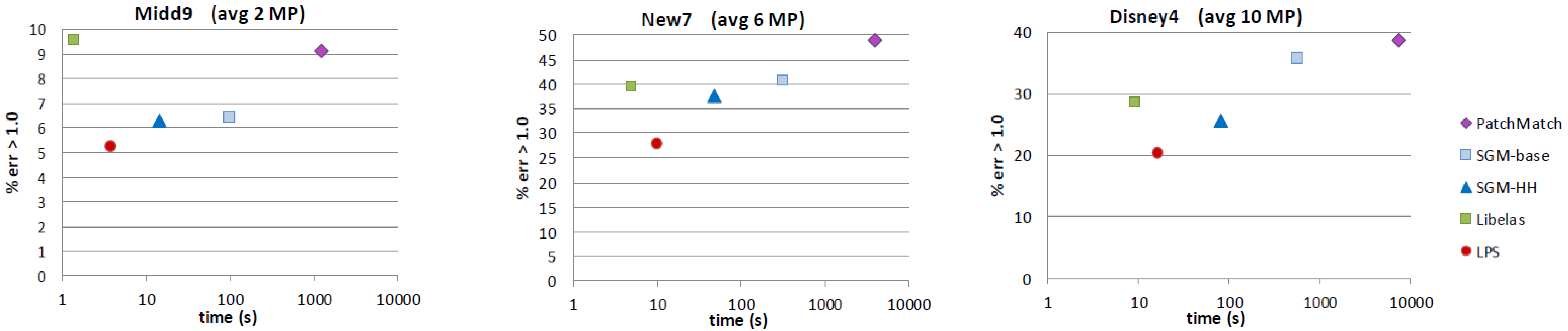




Here, disparities are out of range

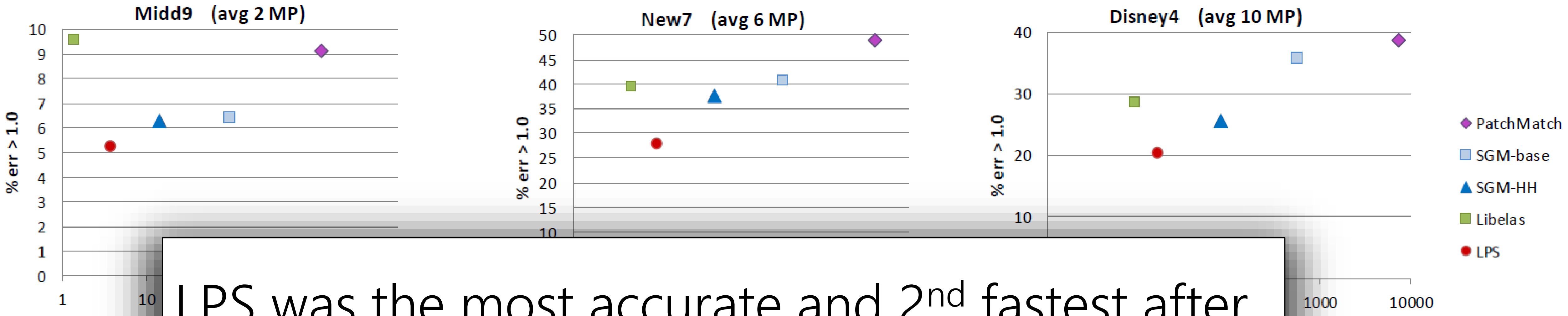
Efficient High-Resolution Stereo Matching using Local Plane Sweeps

[Sinha, Scharstein, Szeliski, CVPR 2014]



Efficient High-Resolution Stereo Matching using Local Plane Sweeps

[Sinha, Scharstein, Szeliski, CVPR 2014]



LPS was the most accurate and 2nd fastest after

- **ELAS: Efficient Large-scale Accurate Stereo**

[Geiger, Roser, Urtasun, ACCV 2010]

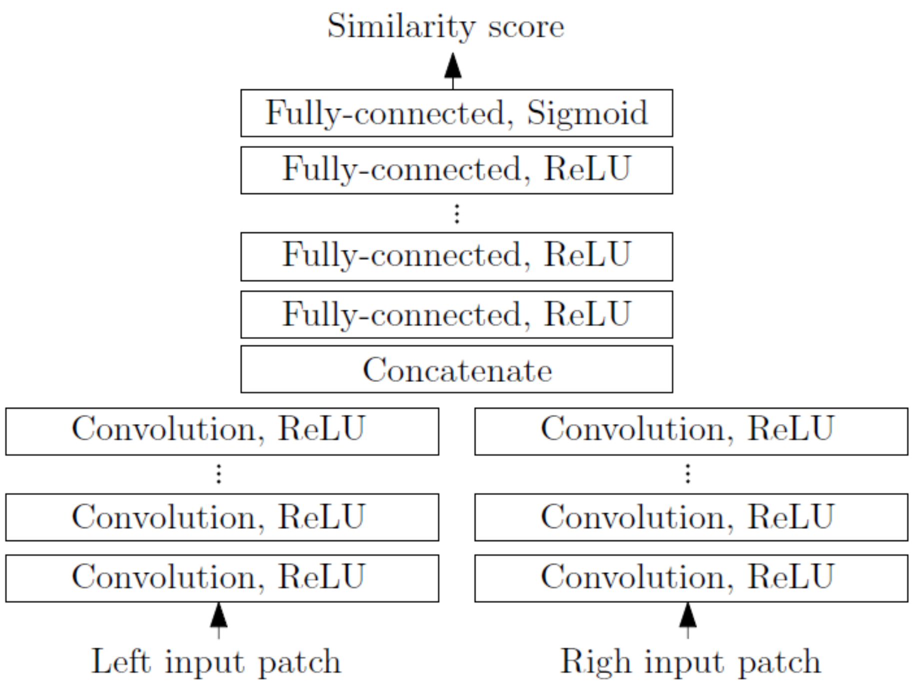
Efficient Stereo Video Processing

Left Camera View

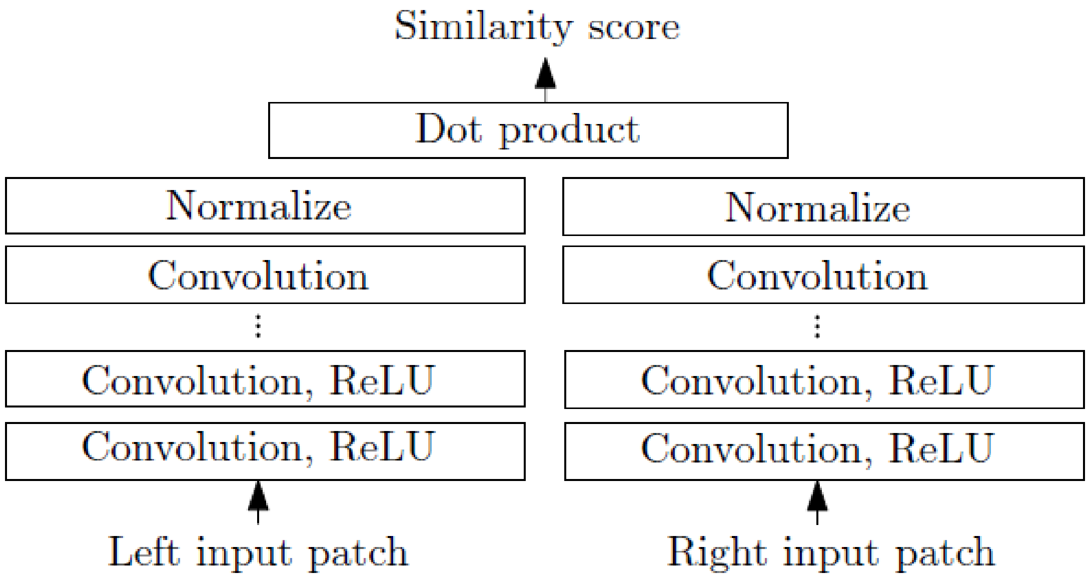
Disparity Map
(Depth)

Detected Moving
Objects

Convolutional Neural Networks for Correspondence



Accurate Architecture



Fast Architecture

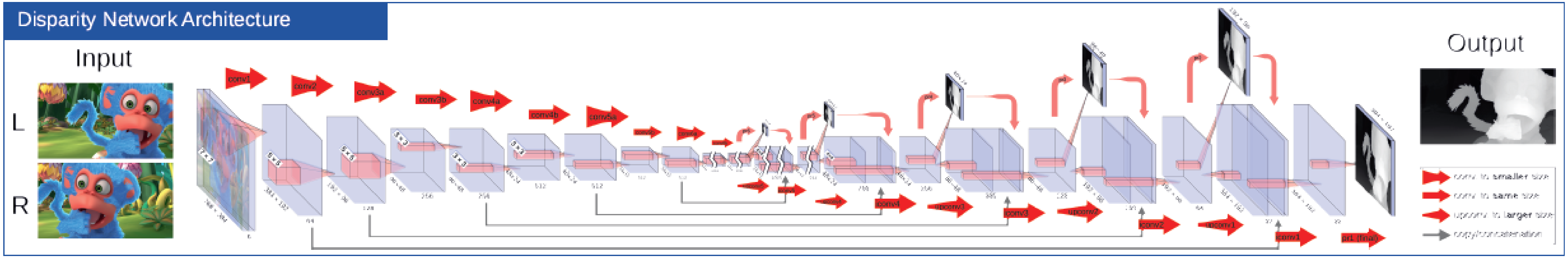
Siamese Networks

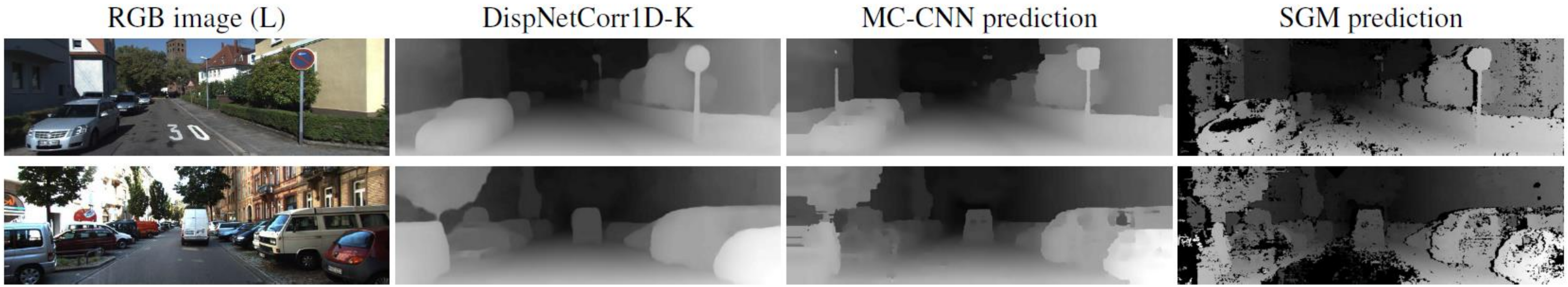
- Local feature descriptors: [Han+ 2015, Zagoruyko+ 2015, Simo-Serra+ 2015]
- Stereo Matching Cost: [Zbontar and Lecun 2015, 2016, Chen+ 2015, Luo+ 2016]

Convolutional Neural Networks for Correspondence

End-to-end deep models:

FlowNet [Dosovitskiy+ 2015], DispNet [Mayer+ 2016]





Overview

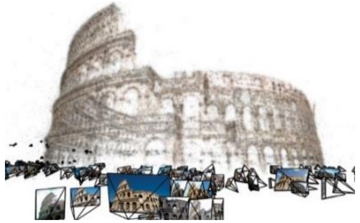
- **Sparse Correspondence and Applications**
  - Place recognition
  - Color Transfer and Enhancing Photos
- **Dense Correspondence Estimation**
  - Stereo Matching on High Resolution Images and Video
- **Joint Correspondence and Cosegmentation**
  - Align images of different but semantically related objects

Semantic Correspondence Estimation

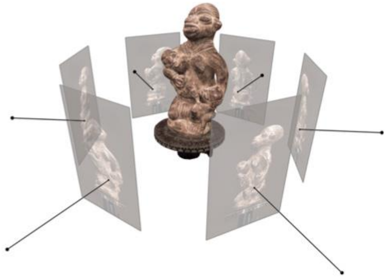
Identical Scene

- Well-studied sub-topics
- Well defined notion of visual similarity

Structure from motion

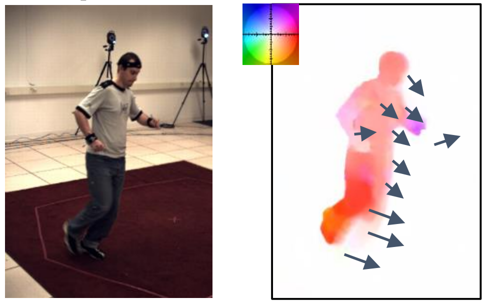


Multiview stereo



Binocular Stereo

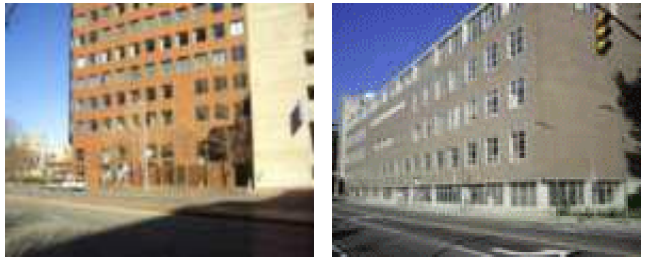
Optical flow



Different but semantically related scenes

- Image appearance could differ a lot
- Much more challenging

SIFT Flow [Liu+ 2008]

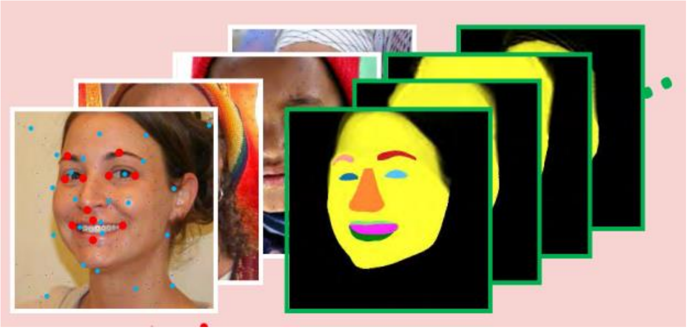


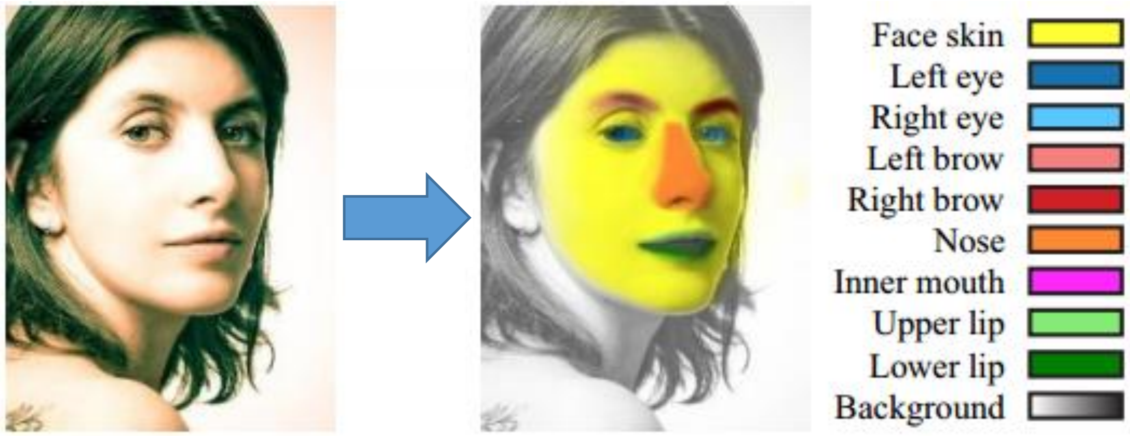
Deformable Spatial Pyramid Matching [Kim+ 2013]

Applications

Label Transfer (Face Parsing) [Smith+ 2013]

Labeled images





Input Output Label Types

Depth Transfer [Karsch+ 2012]

RGB-D database

Query Image Predicted Depth map

Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]

Source   Target   Mask + Flow   Warped Source Image

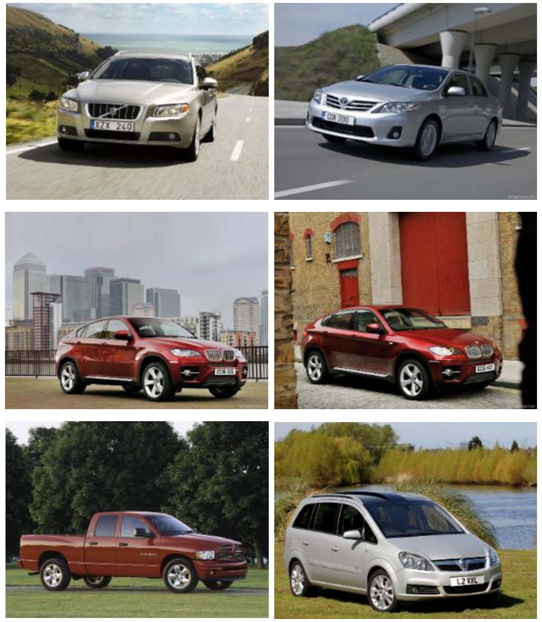
Input

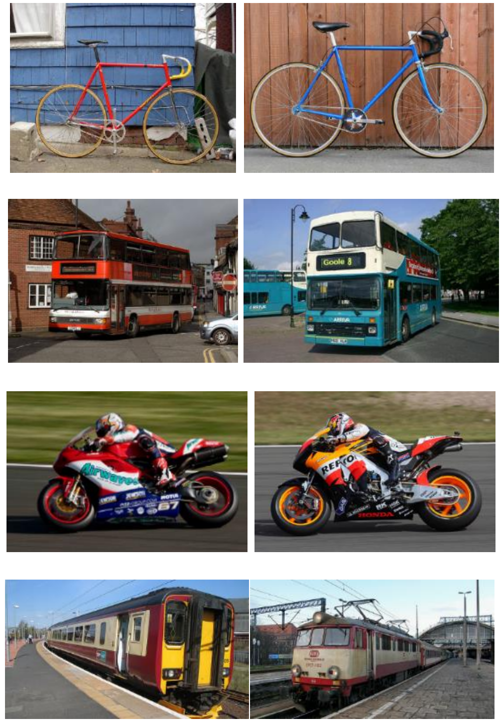
Image pair containing semantically related objects but different instances

Output

Find the common region i.e. foreground (binary) mask and the dense optical flow associated with the common region.

Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]





- Objects from same but unknown category
- Different scene backgrounds
- Visual appearance, object contours, camera viewpoints are dissimilar

Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]

- Model: Hierarchical Layered graph of nested image regions

  (Continuous Label Space)

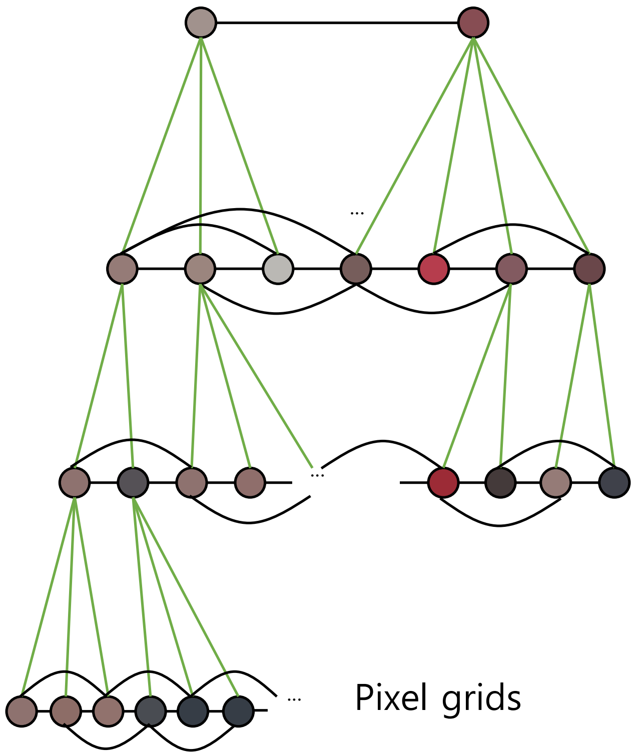
  - binary (segmentation)
  - 2D similarity transform (flow) (4-dof)

  (Spatial regularization)

  - between neighbors
  - between parent-child nodes.

- Energy minimization/Inference

  Local alpha expansions (graph cuts) [Taniai et al. 2014]

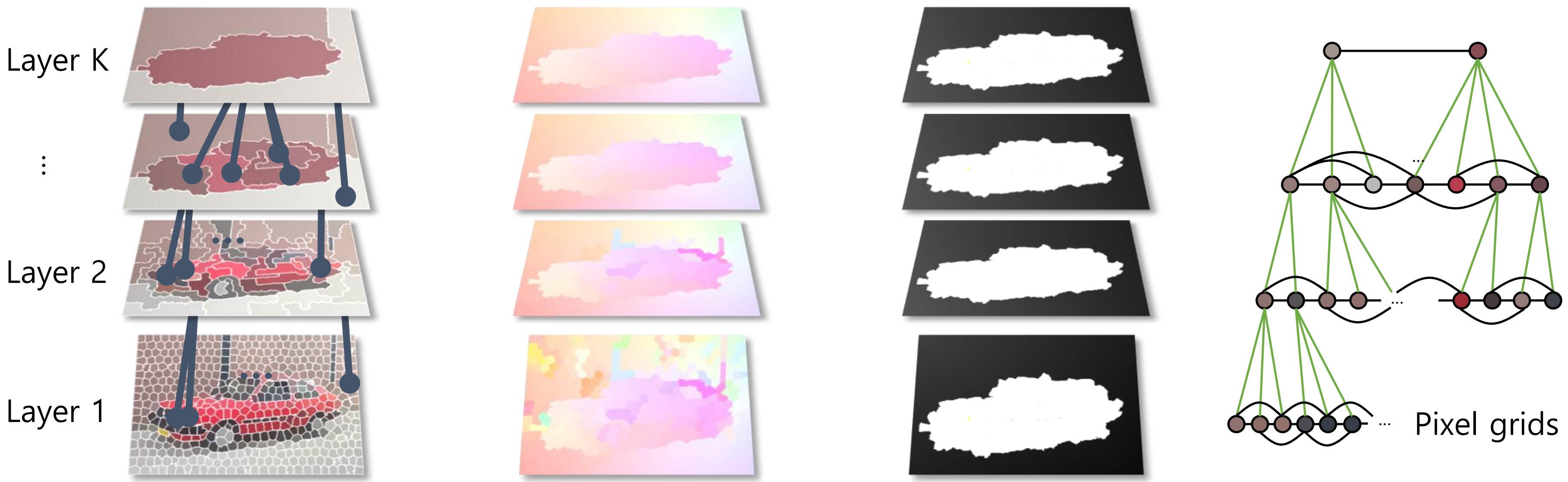


Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]

Microsoft Research Asia Faculty Summit 2016

$$F(G,f,\alpha) = E_{graph}(G) + E_{flow}(f|G) + E_{seg}(\alpha|G) + E_{reg}(f,\alpha|G)$$

Layered graph **Flow term** **Segmentation term** **Smoothness terms**

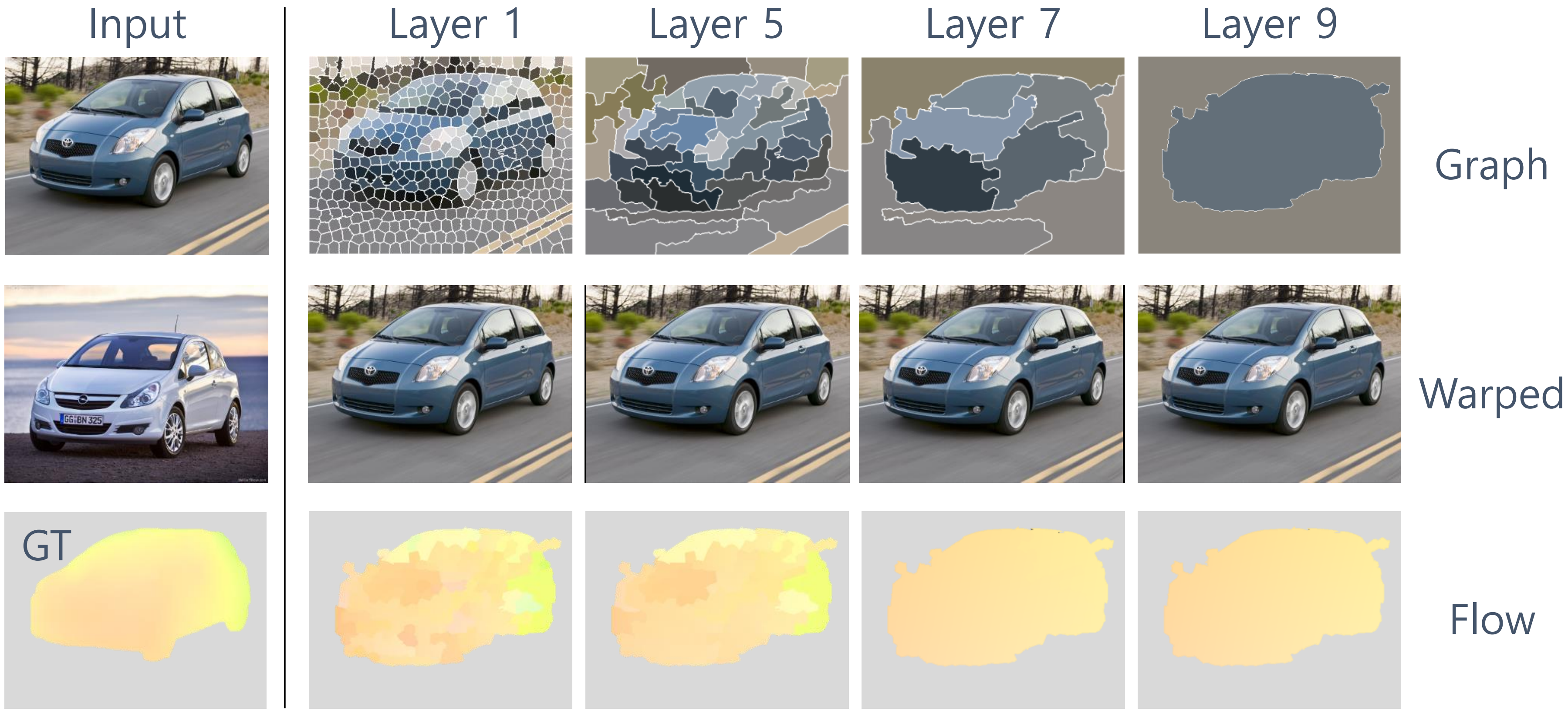


- structure inferred one layer at a time
- Patch matching with HOG descriptors
- FG/BG color likelihoods

- Spatial neighbor edges
- Parent child edges

Microsoft

Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]



Our Hierarchical Model

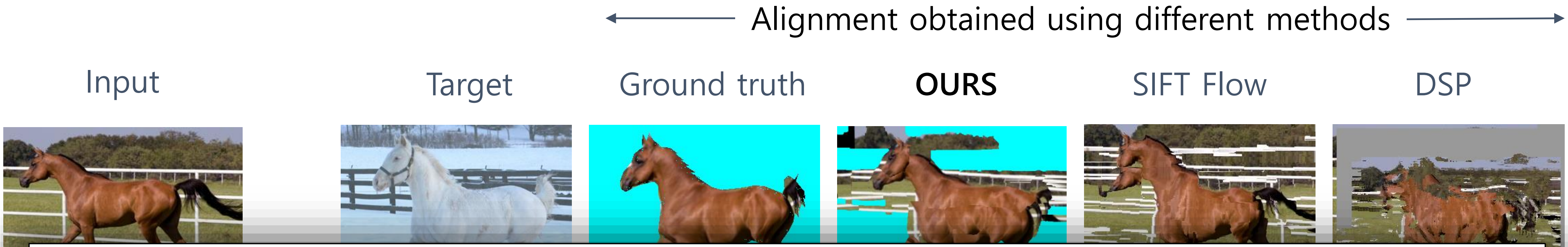
Joint Cosegmentation and Dense Semantic Correspondence

[Taniai, Sinha, Sato, CVPR 2016]

← Alignment obtained using different methods →

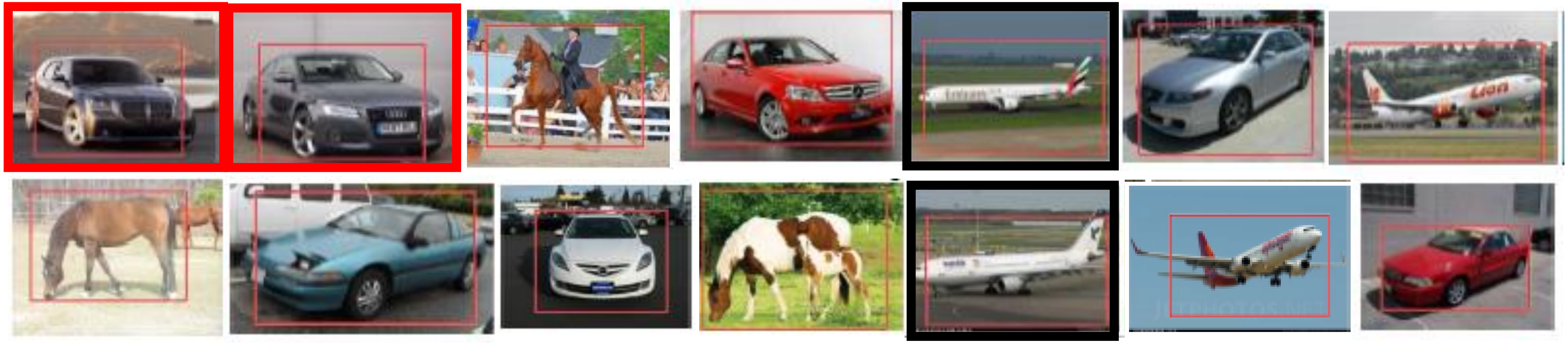
Input | Target | Ground truth | **OURS** | SIFT Flow | DSP

Joint Cosegmentation and Dense Semantic Correspondence [Taniai, Sinha, Sato, CVPR 2016]



- Solving cosegmentation and alignment simultaneously improves accuracy on both tasks.
- Outperforms methods specifically designed for each task

Future Work: Multi-image semantic correspondence



- Unsupervised or weakly supervised setting
- Visual Object Discovery (find the common objects)
    - bootstrap from easy image pairs ?
    - Incremental representation learning

Conclusion

- Sparse Visual Correspondence
  - Self-supervision for feature learning
  - Enables automatic label propagation
- Challenges in Dense Correspondence Estimation
  - High resolution stereo matching, optic flow
  - Efficient stereo video processing
- Semantic Correspondence
  - Unsupervised visual object discovery